Imputation Variance Estimation by Bootstrap Method for the National Ambulatory Medical Care Survey

Yan Li, Carrie Lynch, and Iris Shimizu, National Center for Health Statistics;
Steve Kaufman, National Center for Education Statistics
Yan Li, NCHS, 3311 Toledo Road, Room 3206, Hyattsville, MD, 20782

KEY WORDS: Complex sample survey, Imputation, Variance, Bootstrap

1. Introduction[1]

The National Center for Health Statistics (NCHS) conducts the National Ambulatory Medical Care Survey (NAMCS) to produce annual statistics about ambulatory medical care provided by office based physicians in the U.S. Data from the survey are released to the public in various forms, including micro data in public use files (PUFs). The survey uses a complex sample of visits that are made to physician offices for the purpose of obtaining medical care. Physicians and/or their staff are asked to abstract data about each visit onto a NAMCS patient record form (PRF). As commonly happens in surveys, the NAMCS estimates are subject to item nonresponse – that is, a respondent fails to complete some PRF item and that item's missing data cannot be extracted from data which are supplied for other items on that visit's PRF. Four items included in NAMCS PUFs are imputed when item nonresponse occurs. Three of these items (age, sex, and race) are imputed using a cold deck procedure during data keying and editing operations while the fourth item (time spent with physician) is imputed later with a hot deck procedure. Hence, the NAMCS estimates for those four variables are subject to imputation variance. This paper describes a methodology explored for estimating the imputation variances for variables in the NAMCS PUF. Experimental variance estimates were produced for the specific imputation procedures applied to the 2000 NAMCS.

A rich body of literature has proposed procedures to estimate the variance due to imputation. Among them are the multiple imputation procedure by Rubin (1987), a number of model-assisted estimators of variance proposed by Sarndal (1992), and adjusted jackknife variance estimators for both random and deterministic imputations (Rao 1993; Rao and Shao 1992; Rao and Sitter 1995) under stratified multistage sampling. However, all of these proposed procedures have their own limitations. Multiple imputations may not provide consistent variance estimators and may have poor empirical performance when the sampling design is stratified sampling and/or multistage cluster sampling (Fay 1991, 1993), whereas jackknife cannot be applied to the case where the statistic is nonsmooth. Shao and Sitter (1996) proposed an adjusted bootstrap method. According to them, this method is asymptotically valid irrespective of the sampling design, the imputation method, or the type of statistic used in inference. In the current research, an adjusted bootstrap method is applied to produce NAMCS's first estimates of imputation variance. Readers interested in various bootstrap methods and their applications are referred to a recent review paper by Lahiri (2003).

Section 2 describes the NAMCS survey design. Section 3 discusses the details of procedures used to impute for item nonresponse in NAMCS while section 4 describes the procedures used to derive estimates of the variance due to that imputation. Results are discussed in Section 5 while findings are summarized in Section 6.

2. Survey design and operations

The NAMCS is a national probability sample survey of ambulatory medical encounters in physician's offices. The universe of physicians for the survey consists of non-Federally employed physicians classified by the American Medical Association (AMA) or American Osteopathic Association (AOA) as "office-based, patient care" physicians. Physicians in the specialties of anesthesiology, radiology, and pathology are excluded. Visits to private, nonhospital-based clinics are included, but those that occur in federally operated facilities, industrial health clinic settings, and hospital-based outpatient departments are not. Telephone contacts and non-office visits are also excluded, as well as purely administrative visits (e.g., bill payment, leaving specimens).

The NAMCS uses a multistage probability design. The first stage sample is a probability subsample of the geographic primary sampling units (PSUs) used by the 1985-94 National Health Interview Survey. From the sample PSUs, a stratified sample of physicians is selected from the master files of the AMA and the AOA with strata defined by physician specialty groups. The sampled physicians in each specialty group are randomly distributed among the

[1] The opinions expressed in this paper are those of the authors and not necessarily those of the National Center for Health Statistics.

52 one-week reporting periods in a calendar year so that NAMCS estimates account for seasonal variation. The physicians are asked to complete PRFs for a systematic random sample of visits made to the physician's practice during the physician's reporting week. For the 2000 NAMCS, the sample included 3,000 physicians of which 2,049 were eligible to participate in the survey. About 68 percent of the eligible physicians participated and a total of 27,369 PRFs were completed.

The U.S. Census Bureau collects the NAMCS data and sends it on a flow basis to NCHS's data processing contractor, where the PRF data are edited, coded, and keyed for entry into electronic media. The contractor immediately imputes the missing data for birth date (from which age is derived), sex, and race after keying each PRF's data and before doing computer edits for code ranges and consistencies. If an imputed value then fails consistency checks, the keyer returns to the PRF to determine which value is at fault and can take one of several actions: (a) correct errors in coding or keying of a non-imputed value, (b) enter the exact value for the imputable item, if one is then discovered or extracted from other data in the PRF, or (c) start the imputation over if neither of the other two actions apply. The cycle of imputation and consistency checks is repeated for each PRF until all of its imputed values pass consistency checks. All imputed values are flagged in the data file. During the contractor's processing, a 10 percent systematic random sample of PRFs is independently recoded in every batch and the whole batch is recoded if 5 percent of the sample fails. Data processing details for the 2000 data are included in National Center for Health Statistics (1999).

NCHS completes preparing the data for use and release to the public. First, independent computer edits are conducted and the remaining imputation (for the item "time spent with physician") is done. Next the data are weighted to produce national estimates in four basic steps: (a) inflation by reciprocals of the sampling selection probabilities, (b) adjustment for nonresponse, (c) calibration ratio adjustment (at the physician level), and (d) weight smoothing. Finally, design variables required for variance estimation are added to the weighted data files. In the PUFs the design variables are masked to minimize disclosure risks for respondents while at the same time enabling public data users to approximate the variances for their own statistics. The quality of variance approximations derived from the PUFs was examined by Hing et al (2004). Imputation variances produced in this research were approximated from the PUFs and may differ from those for the NAMCS published variances which are derived with unmasked design variables available in the in-house files.

3. Imputation procedures

One of the four survey items imputed in the NAMCS PUFs is the time spent with physician (TIMEMD). If a physician does not report TIMEMD for a sample visit, the survey item is imputed using a hot deck procedure by selecting a donor from a pool of visits in the current sample. Potential donors are sample visits for which the physician reported TIMEMD and also reported seeing the patient.

The imputation of items in an individual visit record is performed within imputation classes, most of which are defined by the physician's responses to other relevant survey items on the questionnaire. The first imputation class is defined by three survey items: DIAG is the 3-digit ICD-9-CM code for the physician's primary diagnosis; SPEC is the physician's specialty; and REGION[2] is the U.S. geographic region (Northeast, Midwest, South, or West) of the physician's office. The second, third, and fourth imputation classes use a subset of the variables in the first imputation class. They are defined by DIAG and SPEC; DIAG; and SPEC and REGION, respectively. Finally, the fifth imputation class is not defined by relevant survey items. It is one, all inclusive group.

When imputing for missing TIMEMD data, visits for which TIMEMD is missing are referred to as donees. The donees and donors are independently sorted by DIAG, SPEC, and REGION. The donees are imputed sequentially within each class defined by DIAG, SPEC, and REGION. For each donee, a donor is randomly selected without replacement from the pool of donors having the same DIAG, SPEC, and REGION class as the donee, and the donee is assigned the donor's value of TIMEMD. If there are no potential donors in the donee's class, the donee is placed in its second imputation class for later imputation. When the first pass of the data is completed, all missing TIMEMD data will either be imputed in the first imputation classes or the donees

[2] The value of region used in this study is the one recorded in PUF visit records and is the region where the doctor's practice was located at interview time. The original imputation, however, was done using the region value recorded in the physician sampling frame for the physician because that is the only region value available when the original imputation is done for age, sex, and race. The sampling and interview regions will differ for sample doctors who moved between the time they were added to the sampling frame and the time they were interviewed. Relatively few sample physicians move between regions so the effect on imputation variances due to substituting interview for sampling region is probably negligible.

are placed in their second imputation classes. A similar imputation process is repeated successively within the second, third, fourth, and fifth imputation classes, if necessary, until all missing TIMEMD data items are filled.

The NAMCS PUFs also contain imputed data for age, sex, and race of the patient when the physician does not report those data. These three missing data items are imputed simultaneously using a cold deck procedure which selects a donor from a pool of visits in the previous year's sample. Potential donors are sample visits for which the physician reported the age, sex, and race of the patient. The donors came from the previous year's sample because age, sex, and race imputations are done in tandem with data keying operations, i.e., missing values are filled for a given record prior to keying data for the next record. Consequently, the current year's sample can not be used because it does not contain clean, edited data for all records.

The imputation classes for age, sex, and race are identical to those used for imputing TIMEMD. However, the imputation procedure is slightly different. The first difference is donors were mistakenly selected with replacement for the 2000 NAMCS. In other survey years, donors were selected without replacement. The second difference is consistency checks were performed to ensure that the imputed age and sex values agree with the data that were reported in the donee's PRF. If the donor passes consistency checks, then the donee's missing data items are filled with the donor's values. Otherwise, the donor is rejected and dropped from the donor pool. In addition, the donor pool is re-randomized, another donor is randomly selected, and consistency checks are performed. This process is repeated until a donor passes consistency checks. Moreover, the donor pool is restored before imputing the next donee.

It was not possible to determine the exact order in which donees' missing age, sex, or race were imputed in the 2000 NAMCS PUF because imputation and data keying were done in tandem. However, it is believed that the order described in this section should not affect the magnitude of the variances because the donors were selected with replacement.

4. Bootstrap procedure

Prior to estimating sampling and imputation variances using the bootstrap procedure, the replicate weights required for the procedure were computed as follows. Suppose that n_h PSUs were drawn with replacement from stratum h. In each stratum a simple random subsample of m_h PSUs were drawn with replacement from the n_h PSUs. Shao and Tu (1995) contains a collection of investigations into the best choice of m_h for various n_h. The selected value of m_h was that one proposed by McCarthy and Snowden (1985) which simplifies to $m_h = n_h - 1$ in the current study. Finally, the replicate weight for sample visit k within PSU i from stratum h in subsample b was computed by

$$w_{bhik} = w_{hik} \frac{n_h}{m_h} M_{bhi},$$

where w_{hik} represents the survey weight including the adjustments and smoothing as described in Section 2 and M_{bhi} denotes the number of times that PSU i was selected in subsample b.

The Shao and Sitter (1996) method was implemented in the following three steps. In the first step, 99 bootstrap subsamples were constructed and the potential donors were flagged. (A larger number of samples would be desirable to increase the stability of the resulting variance estimates, but the number of samples was arbitrarily set at 99 to conserve resources.) Bootstrap subsample b contained sample visits for which $w_{bhik} \neq 0$. For imputing TIMEMD in bootstrap subsample b, the potential donors were the sample visits in bootstrap subsample b for which the physician reported TIMEMD and also reported seeing the patient in the 2000 NAMCS. For imputing age, sex, and race in each bootstrap subsample, the potential donors were sample visits for which the physician reported sex, age, and race of the patient in the 1999 NAMCS.

In the second step, the imputed bootstrap subsamples were created. For imputed bootstrap subsample b, the missing data for imputable items (TIMEMD, age, sex, or race) were imputed using the imputation procedures presented in Section 3 and the donors in bootstrap subsample b as described in step one above. The sample visits with reported data remained in the imputed bootstrap subsample.

Regardless of which survey item was missing data, the seed for generating the random numbers for selecting donors was changed so that the same seed was not used in any two replicates. In addition, a different seed was also used within each imputation class.

For the third step, consider the following notation. Let B denote the number of imputed bootstrap subsamples; $\hat{X}$ represent the characteristic estimate computed using the entire sample weighted by survey weight w_{hik} and $\hat{X}_b$ denote the same statistic computed using imputed bootstrap subsample b weighted by replicate weight w_{bhik}. The variance which accounts for both sampling and imputation variation in estimate $\hat{X}$ was approximated by

$$v(\hat{X}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{X}_b - \hat{X})^2 .$$

Bootstrap variances were also estimated when variation due to imputation was excluded. The variance formula presented above was used except $\hat{X}_b$ represents the characteristic estimate computed from each subsample *b* using the original imputed values from the 2000 NAMCS PUF instead of re-imputed values.

In the variance formula presented above, the deviation is around the full sample estimate while, in Shao and Sitter's variance estimator, the deviation is around the average of the imputed bootstrap subsample estimates. Zhang et al (1998) showed that the bias in the variance estimator presented above can be corrected by multiplying the variance estimator by an adjustment factor of *(B-1)/B*. For this research, the square root of the adjustment factor is $\sqrt{98/99} = 0.99$. Because the factor is close to 1, it was not applied.

5. Results

The characteristic estimates considered in this study for TIMEMD, age, sex, and race were restricted to those published in NCHS' Advance Data Report for the 2000 NAMCS (Cherry and Woodwell 2002).

Ratios of bootstrap standard errors (SEs) were used to evaluate the effect of imputation on variances. The numerators of these ratios are the SEs estimated by re-imputing the missing values across the bootstrap subsamples while the denominators are the corresponding SEs estimated by using the original imputed values in all bootstrap subsamples. The ratio numerators include variation due to imputation while the denominators do not.

Figures 1 and 2 show the impact of imputation of TIMEMD has on the bootstrap SEs. Figure 3 shows the effect on the bootstrap SEs by imputation of sex, age, and race. The bootstrap SEs discussed in this section are estimates of SEs and should not be considered true SEs.

Figure 1 shows the bootstrap SE ratios for the estimated number of office visits by TIMEMD. It can be noticed that the SE ratios for all the time intervals are greater than one and the largest two SE ratios come from the time intervals of "Over 60 minutes" and "1-5 minutes." It can be calculated that among the sample visits for which physicians reported TIMEMD, fewer than 700 visits fell in each of these two time intervals while over 2,000 visits fell in each of the other time intervals.

Figure 2 shows the distribution of bootstrap SE ratios for the estimated mean time spent with physician for 14 physician specialty groups. It can be observed that 3 specialty groups have bootstrap SE ratio less than 1.06, the other 11 specialty groups have relatively larger SE ratios and among them, 5 specialty groups have bootstrap SEs inflated 16 to 31 percent due to imputation. The relationship between the bootstrap SE ratios and imputation rates for the estimated mean time spent with physician for 14 specialty groups was investigated. The Pearson Correlation Coefficient was $\hat{\rho} = 0.78$, which is positively significant.

Figure 3 shows the distribution of bootstrap SE ratios by ratio range for estimated numbers of visits by selected patient demographic characteristics and major reason for visit. For estimated visits by sex and major reason for visit, all 12 of the corresponding SE ratios were less than 1.03. The imputation rate for sex is only 0.68 percent. For estimated visits by age and major reason for visit, most the corresponding SE ratios are less than 1.03 and none are greater than 1.05. The imputation rate for age is 3 percent and is larger than that for sex. For estimated visits by race and major reason for visit, imputation increases the bootstrap SEs by 6 to 13 percent for 5 of the 18 estimated numbers of visits. Such increases for estimates by race can be expected because the imputation rate for race is 18 percent, which is much larger than that for sex or age. Bootstrap SE ratios were affected by imputation rates.

6. Summary

This paper studied the impact of imputation on the standard error estimates by using a bootstrap method for the 2000 National Ambulatory Medical Care Survey (NAMCS) public use file (PUF). Four items included in NAMCS PUFs are imputed when item non-response occurs. These are "time spent with the physician," sex, age, and race. The empirical results in the current study demonstrate increases in variance due to imputation for selected published estimates involving the imputed variables.

It was observed that variance increases caused by imputation tend to vary with imputation rates and inversely with the number of sample visits with reported (not imputed) data contributing to each estimate. That is, large variance increases (6 to 31 percent) were observed for some statistics afflicted with high imputation rates and/or based on "low" numbers of unweighted visits having reported data. However, for the majority of study statistics (65 out of 86), the variation due to imputation appears small (less than 6 percent) and, thus, suggests that imputation has only minimal impact on analysis involving those statistics.

The results of the current study are subject to limitations. First, the results may be unique to the 2000 NAMCS because that is the year when donor

selection for imputation of age, sex, and race was done with replacement instead of without replacement as is done in other data years. Second, the observed results may differ from those that would be computed using the unmasked design variables instead of the PUF masked design variables used in the current study.

More study is required before detailed advice and/ or procedures can be provided that will enable data users to account for variation due to imputation in their analysis of NAMCS data. In addition to evaluating the potential limitations mentioned in the above, the bootstrap methods used in this study should be compared with other imputation variance procedures and evaluated in terms of both results and ease of application to NAMCS data. In the mean time, this study's results reinforce the need for analysts to use conservative analytic techniques in the presence of masked design variables and imputed data.

References

Cherry, D. K. and Woodwell, D. A. (2002), "National Ambulatory Medical Care Survey: 2000 Summary," Advance Data from Vital and Health Statistics, No. 328, Hyattsville, Maryland: National Center for Health Statistics.

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance," in *Proceedings of Seventh Annual Research Conference*, U.S. Bureau of the Census, pp. 429-440.

——(1993), "Valid Inferences from Imputed Survey Data," in *Proceedings of Survey Research Methods Section, American Statistical Association*, pp. 41-48.

Hing, E., Gousen, S., Shimizu, I., and Burt, C. (2004), "Guide to Using Masked Design Variables to Estimate Standard Errors in Public Use Files of the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey," *Inquiry,* 40(4):416-415.

Lahiri, P. (2003), "On the Impact of Bootstrap in Survey Sampling and Small Area Estimation," *Statistical Science*, 18, 199-210.

McCarthy, P. J. and Snowden, C. B. (1985), "The Bootstrap and Finite Population Sampling," National Center for Health Statistics, *Vital Health Stat,* 2(95).

National Center for Health Statistics (1999), "General Processing Instructions, Imputations, and Consistency Checks," in National Ambulatory Medical Care Survey/National Hospital Ambulatory Medical Care Survey 1999 Coding Requirements, Internal document, Hyattsville, MD: Author.

Rao, J. N. K. (1993), "Linearization Variance Estimators under Imputation for Missing Data," Technical report, Carleton University, Laboratory for Research in Statistics and Probability.

Rao, J. N. K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika*, 79, 811-822.

Rao, J. N. K. and Sitter, R. R. (1995), "Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data," *Biometrika*, 82, 453-460.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: J. Wiley & Sons, Inc.

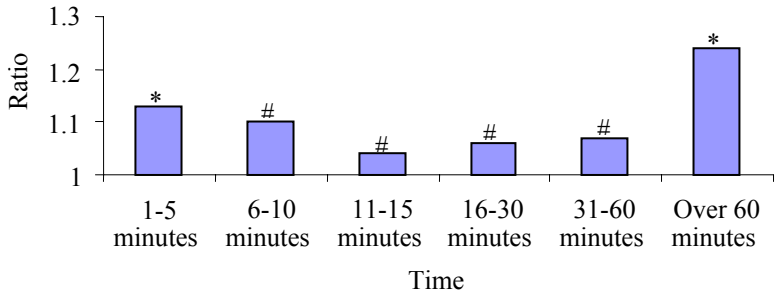
Sarndal, C. E. (1992), "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used," *Survey Methodology*, 18(2): 241-252.

Shao, J. and Sitter, R. R. (1996), "Bootstrap for Imputed Survey Data," *Journal of American Statistical Association*, 91, 1278-1288.

Shao, J. and Tu, D. (1995), *The Jacknife and Bootstrap*, New York: Springer-Verlag.

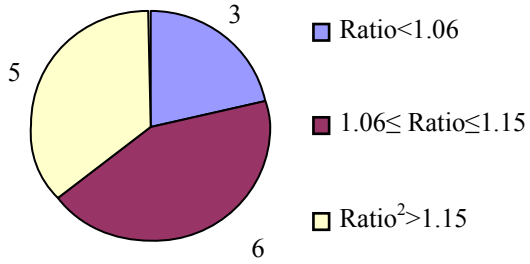
Zhang, F., Brick, M., Kaufman, S., and Walter, E. (1998), "Variance Estimation of Imputed Survey Data," Working Paper No. 98-14, Washington, D.C.: National Center for Education Statistics.

Figure 1. Bootstrap standard error ratios for number of office visits by time spent with physician



* Fewer than 700 sample visits
# Greater than 2,000 sample visits

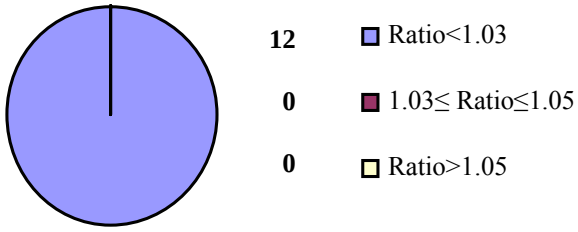
Figure 2. Distribution of bootstrap standard error ratios for estimates of mean time spent with physician[1]



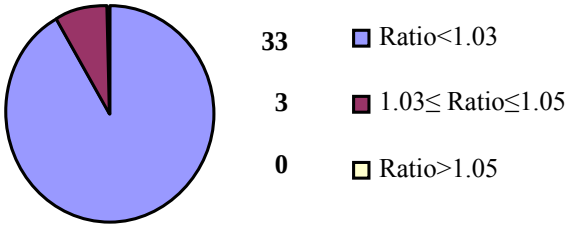
[1] Mean time spent with physician was estimated for 14 physician specialty groups.
[2] The maximum of the bootstrap standard error ratios is 1.31.

Figure 3. Distribution of bootstrap standard error ratios for estimated visits by selected visit characteristics

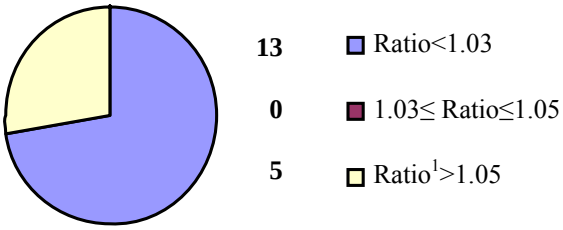
Sex and major reason for visit



Age and major reason for visit



Race and major reason for visit



[1] The maximum of the bootstrap standard error ratios is 1.13.